

Curso avanzado sobre grandes modelos de lenguaje: fundamentos y aplicaciones
Dr. Trappolini Giovanni (Profesor Visitante FCEN-UBA) con la colaboración del Dr. Pablo Barenbaum (Profesor Interino Adjunto)

Programa:

Este curso sobre grandes modelos de lenguaje (LLMs) tiene como objetivo dar a los estudiantes de una comprensión integral de estos modelos, equilibrando visión teórica y aplicaciones prácticas.

El objetivo principal es desarrollar el dominio de los conceptos fundamentales que rodean a los LLMs, permitiendo a los estudiantes reconocer los mecanismos y principios subyacentes que hacen que estos modelos sean simultáneamente poderosos y sutiles.

Desde un punto de vista teórico, los estudiantes se sumergirán en la arquitectura de los LLMs, las metodologías de entrenamiento y los fundamentos matemáticos.

Desde el punto de vista práctico, los estudiantes aprenderán a implementar arquitecturas básicas de LLM desde cero, proporcionándoles una experiencia directa de los desafíos y sutilezas involucrados.

Aprovechando bibliotecas de última generación, los estudiantes practicarán con LLMs de vanguardia, entendiendo cómo utilizarlos de manera efectiva para tareas diversas.

Al completar este curso, los estudiantes estarán equipados para utilizar estos modelos de manera independiente en sus actividades.

Temario:

- Fundamentos del aprendizaje profundo. Regresión lineal/logística simple. Redes lineales profundas.
- Mecanismos de atención y el problema de la traducción automática.
- Arquitecturas basadas en transformadores (en particular BERT y GPT). Uso de estos modelos en un entorno de aprendizaje por transferencia.
- Prompting, capacidades de transformers basados en pocos o ningún "shot". Formas eficientes para afinar modelos grandes con pocos recursos computacionales (PEFT, LORA).
- Aprendizaje por refuerzos, retroalimentación humana y optimización de preferencia directa.
- Aplicaciones de modelos de lenguaje grande. Arquitecturas basadas en transformadores en dominios multimodales.
- Los límites de los LLM y su uso en contextos similares a RAG (Recuperación de Información Neuronal).
- Introducción a PyTorch, manipulación de tensores, manejo de datos, implementación de modelos básicos.
- Implementación del bloque de atención básico (un solo cabezal).
- Introducción a la librería Hugging Face. Implementación de tareas posteriores (por ejemplo, respuesta a preguntas).
- Ajuste fino eficiente de modelos grandes con Hugging Face y Peft. Comparación con implementaciones de pocos "shots".
- Implementación de un sistema RAG (Hugging Face, LangChain). Implementaciones de transformadores en el dominio multimodal para tareas posteriores (ViT, VisualBERT).

Bibliográfia:

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
- Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer, 2009.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474, 2020.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837, 2022.